

The Sensitive Artificial Listener: an induction technique for generating emotionally coloured conversation

Ellen Douglas-Cowie¹, Roddy Cowie¹, Cate Cox¹, Noam Amir², Dirk Heylen³

Queen's University Belfast¹, Tel Aviv University², University of Twente³

e-mail: e.douglas-cowie@qub.ac.uk, r.cowie@qub.ac.uk, c.cox@qub.ac.uk, noama@post.tau.ac.il, heylen@cs.utwente.nl

Abstract

The aim of the paper is to document and share an induction technique (The Sensitive Artificial Listener) that generates data that can be both tractable and reasonably naturalistic. The technique focuses on conversation between a human and an agent that either is or appears to be a machine. It is designed to capture a broad spectrum of emotional states, expressed in 'emotionally coloured discourse' of the type likely to be displayed in everyday conversation. The technique is based on the observation that it is possible for two people to have a conversation in which one pays little or no attention to the meaning of what the other says, and chooses responses on the basis of superficial cues. In SAL, system responses take the form of a repertoire of stock phrases keyed to the emotional colouring of what the user says. The technique has been used to collect data of sufficient quantity and quality to train machine recognition systems.

1 Introduction

It is a difficult problem to generate recordings of emotionally coloured conversation data that are reasonably natural, but still suitable for machine analysis. This paper describes an induction method that generates data which has been successfully used in a machine learning environment. The technique is called the Sensitive Artificial Listener Technique, developed at Queen's University Belfast. The aim is to document this tool and share it with the research community.

There have been several published descriptions of analyses that use data from SAL exercises, but there is no generally available description of the technique itself. This paper remedies that omission.

1.1 Background and Context

It has become clear that for different reasons, emotion-oriented computing cannot rely either on data from actors or on fully naturalistic recordings. As a result, there is great interest in data generated by techniques designed to elicit emotion deliberately. This type of approach produces data that can be both tractable and reasonably naturalistic. Many induction techniques are in use in the machine learning context, such as computer games (Bechara, Damasio, Damasio & Anderson 1994; van Reekum et al 2004; Wang and Marsella 2006) or tasks involving computers (Batliner, Fischer et al, 2003; Batliner, Hacker et al. 2003; Aubergé, Audibert & Rilliard 2004) and sometimes tasks involving human-human interaction (Bachorowski & Owren 1995; Abassi et al 2007; Martin et al. 2006).

The Sensitive Artificial Listener is a specific type of induction technique that focuses on conversation between a human and an agent that either is or appears to be a machine. It is designed to capture a broad spectrum of emotional states, expressed in 'emotionally coloured discourse' of the type likely to be displayed in everyday conversation.

It is a challenge to collect records of human-machine conversation, because machines are not actually able to

carry out conversations. However, there are obvious reasons to try, since it seems very likely that human-machine interactions will differ from human-human interactions in significant ways. Not the least of these is that for the foreseeable future, human-machine interactions will break down in ways that human-human interactions do not, and it is important to have ways of recognising the signs of breakdown.

2 The SAL technique

2.1 The basic context

The Sensitive Artificial Listener technique (SAL for short) is based on the observation that it is possible for two people to have a conversation in which one pays little or no attention to the meaning of what the other says, and chooses responses on the basis of superficial cues. The point was made long ago by the ELIZA scenario (Weizenbaum 1996). In the SAL technique, system responses are keyed to the emotional colouring of what the user says, rather than (as in Eliza) words or phrases. The versions used so far have used Wizard of Oz techniques where a human operator follows a script that specifies possible responses. Because the aim is to evoke emotionally coloured responses, the statements are stock phrases chosen to evoke strong reactions in the listener. In current versions, the SAL operator chooses which statement to use at any given time from a menu that is organised to simulate four personalities – Poppy (who aims to make people happy), Obadiah (who aims to make people gloomy), Spike (who aims to make people angry) and Prudence (who aims to make people pragmatic). Users choose at any time which 'personality' they want to talk to. The response that is chosen will depend on the 'personality' that is active and the user's state. The combination creates an environment rich enough to provoke exchanges that are extended, and quite highly coloured emotionally.

2.2 The SAL structure

The four characters are equipped with a set of characteristic responses encouraging the user into responding in differing emotional states. The SAL has no intelligence, only prespecified stock responses.

Speaker: Poppy

User is: negative active

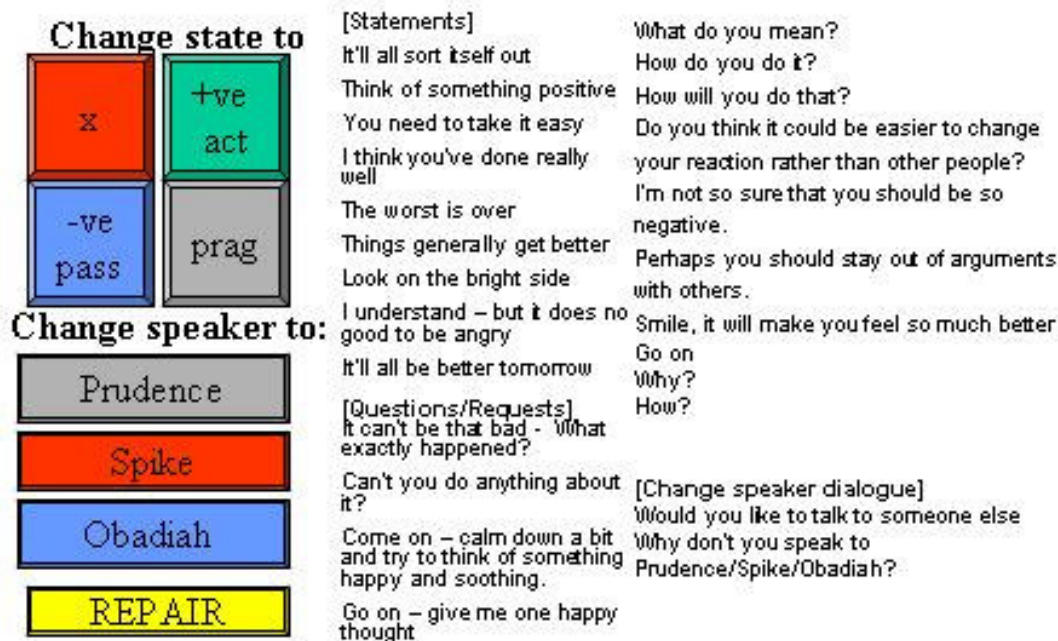


FIGURE 1: Structure governing interaction in SAL

The scripts for the characters were developed, tested and refined in an iterative way. Each character has a number of different types of script depending on the emotional state of the user. So, for example, Poppy has a script for the user in each of four emotional states - a positive active state, a negative active state, a pragmatic state, a negative passive state. There are also script types relevant to the part of the conversation (beginning, main part) or structural state of the conversation (repair script). Each script type has within it a range of statements and questions. They cannot be context specific as there is no 'intelligence' involved. Figure 1 illustrates the type of structure that governs an interaction between one of the personalities of SAL (taken from an early version of SAL).

There have been different versions of SAL moving from an early Wizard of Oz version where the scripts for each personality/character are read by an experimenter who used different tones of voice for the four characters (SAL 0) to a more sophisticated version developed in conjunction with the University of Twente where the phrases are pre-recorded and the experimenter selects phrases from a menu (SAL 1). A fully automated version is currently being developed under the SEMAINE project (<http://www.semaine-project.eu/>). The original version of SAL was in English and was successful enough for versions to be developed in Hebrew (at Tel Aviv University) and Greek (at National

Technical University of Athens, ICCS) with adjustments to suit cultural norms and expectations.

2.2 User experience

SAL has been described as an emotional gym. It does not manipulate users' emotions: that would need far more sophistication. Rather, it gives them prompts to which they can react emotionally if they choose to. It is easy to build up quite a high level of involvement during a sequence of exchanges on an emotive topic. That may be partly because SAL does not inhibit emotional expression by introducing different subjects or perspectives. Various factors lead engagement to break down eventually. SAL responses may simply be too ridiculous for the user to accept; they may become too repetitious; the user may become hopelessly frustrated with SAL's inability to answer questions. Nevertheless, experienced users in particular can easily sustain quite protracted conversations with the system, on the order of half an hour. It appears that listeners learn to use the system, which means that longitudinal use by small numbers has some advantages over occasional use by many.

3 Data

The SAL scenario has been used successfully in three major EU projects (ERMIS, HUMAINE and SEMAINE) to generate large amounts of data that has been labelled and used in a machine learning context.

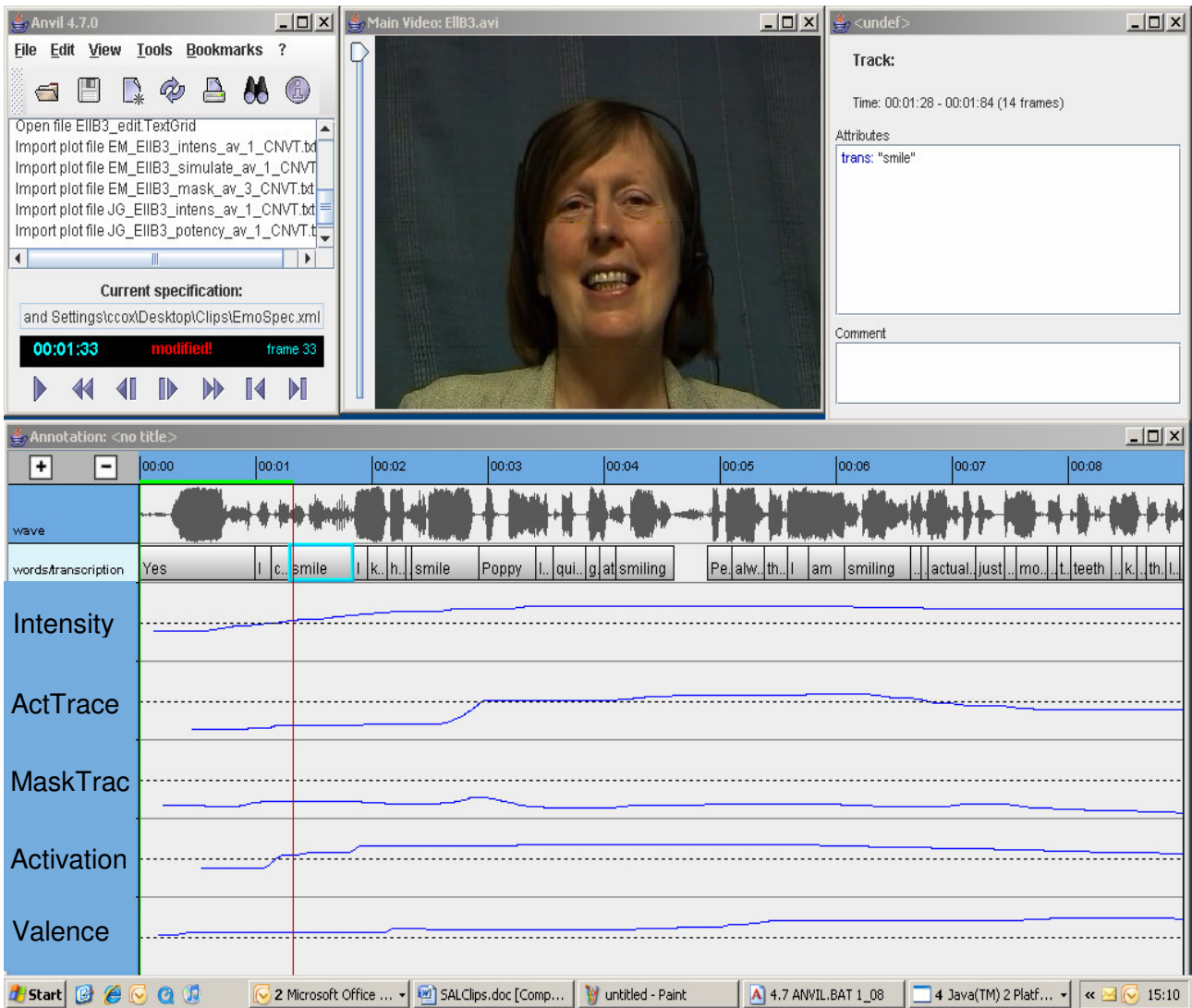


Figure 2: Labelled SAL data from the HUMAINE Database

The data generated is rich in facial and non verbal signals (e.g. aspects of pitch, spectral characteristics, timing), and shows a considerable range of emotions and emotional intensities.

Data was collected using the SAL 0 version from 20 users, 10 male and 10 female. In total 105 minutes of footage was collected. This has been segmented into files and is in avi and mpeg format. There are accompanying files of what was said. SAL 1 was used to collect data from four users, each recorded for two sessions, each of approximately 30 minutes. The data is segmented into files in avi and mpeg format and four raters have labelled the data using the dimensional FEELtrace tool (see Cowie et al. 2000). This gives labels on two dimensions related to emotion (activation and evaluation), and produces traces of how a user's emotional state is perceived over time. A substantial body of data from SAL 1 has also been labelled in a more detailed way as part of the HUMAINE Database (www.emotion-research.net/download/pilot-db/). Data from SAL 1 is releasable under an agreement governing the use of the data. The Hebrew version has undergone

a number of translations and data has now been collected from 5 users, totalling 2.5 hours.

Figure 2 illustrates the kind of data that is produced. It presents a SAL 1 sequence labelled as part of the HUMAINE Database. The emotionality in the user's facial expression is evident. This is borne out by the accompanying traces from a rater for emotion intensity and activation – first and fourth trace lines respectively. The point at which the shot of the face is taken (marked by the vertical red line) corresponds to a rise in the emotional intensity and degree of activation perceived by the rater. The second and third trace lines respectively indicate the degree to which the rater perceives the user to be acting or masking her emotion. The pattern of the ActTrace line indicates a low level of perceived acting at the start rising to absence of acting as the intensity of the emotion rises, indicating the naturalness of the emotion generated.

It is beyond the scope of this paper to describe the statistical properties of the ratings, but Figure 3 summarises some key points. It shows ratings of

valence (x axis) and activation (y axis) in the circular FEELtrace space from two raters. The data covers most of the space except strong negative emotion. The raters agree on the broad pattern, but one (data points in white) is more conservative. Ensuring acceptable consistency is too complex an issue to address here.

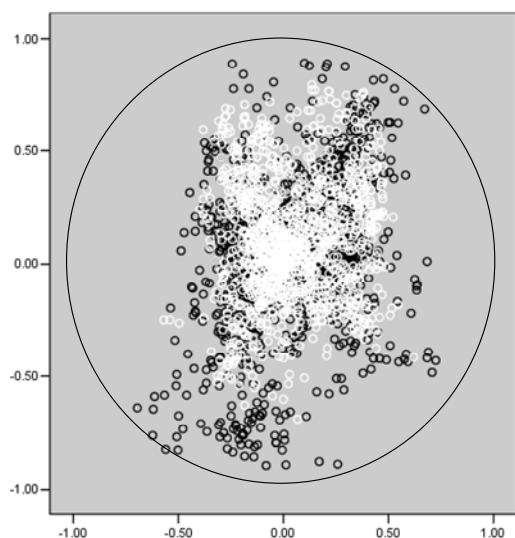


Figure 3: Emotional spread of SAL 1 data

The SAL data that is already available is of sufficient quantity and quality to train machine recognition systems. Published reports of research using the material include Ioannou et al. (2005) and Fragopanagos & Taylor (2005). More recent research reports very high recognition rates when the multimodal character of the data is exploited (Kollias et al. 2008).

4 Conclusion

The success of SAL has led to a new EU funded project called SEMAINE (<http://www.semaine-project.eu/>) which aims to build an automatic human-computer conversation system based on SAL. It will identify the user's emotional state itself, using evidence from face, upper body, voice, and key words. Its range of replies will include some ELIZA-like use of key words extracted from the user's speech. Its own speech will be synthesised, not recorded, and express its emotional stance towards the user. That stance will also be expressed through a graphical display of the 'listener's' face and shoulders. While the user is speaking, the 'listener' will also use vocalisations, facial expressions, and gestures (e.g. nodding) to signal its stance and prompt the speaker to continue or break.

The point of the project is that SAL provides a context in which sustained emotionally coloured human-machine interaction seems to be achievable. Hence, it provides a testbed where it is possible to develop the 'soft skills' needed to sustain such interactions.

Acknowledgement The research reported here has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 211486 (SEMAINE)

5 References

- Abassi, A.R., Uno, T., Dailey, M., Afzulpurkar, N.V. (2007) Towards knowledge-based affective interaction: situational interpretation of affect. In A. Paiva, R. Prada and R. Picard (eds) *Affective Computing and Intelligent Interaction*, Lisbon, September 2007. Berlin: Springer LNCS pp 452-463.
- Aubergé, V., Audibert, N., and Rilliard, A. (2004) E-Wiz: A Trapper Protocol for Hunting the Expressive Speech Corpora in Lab. 4th LREC, 179-182, 2004.
- Bachorowski, J. A. (1999) Vocal expression and perception of emotion. *Current Directions in Psychol Science* 8(2), 53-57.
- Batliner, A., Hacker, C., Steidl, S., Noth, E., and Haas, J. (2003) User States, User Strategies, and System Performance: How to Match the One with the Other. In Proc. ISCA workshop on error handling in spoken dialogue systems, pages 5-10, Chateau d'Oex. ISCA.
- Batliner, A., Fischer, K., Huber, R., Spilker, J., & Noth, E. (2003) How to find trouble in communication. *Speech Communication* 40, 117-143.
- Bechara, A., Damasio, A., Damasio, H., & Anderson, S. (1994) Insensitivity to future consequences following damage to human prefrontal cortex. *Cognition* 50, 7-15.
- Cowie, R., Douglas-Cowie, E., Savvidou, S., McMahon, E., Sawey M., Schroeder, M., 2000 'FEELTRACE': An instrument for recording perceived emotion in real time. In: Proc. ISCA ITRW on Speech and Emotion, Newcastle, N. Ireland, September 5-7, 2000 pp. 19-24.
- Fragopanagos, N & Taylor, J. (2005) Emotion recognition in human-computer interaction. *Neural Networks* 18, 389-405.
- Ioannou, S., Raouzaoui, A., Tzouvaras, V., Mailis T., Karpouzis, K., Kollias, S. (2005) Emotion recognition through facial expression analysis based on a neurofuzzy Network. *Neural Networks* 18, 423-435
- Kollias, S. et al. (2008) HUMAINE IST 507422 Final report for WP4 (www.emotion-research.net)
- Martin, J. C., Devillers, L., Zara, A., Maffiolo, V. and LeChenadec, G. (2006) The EmoTABOU Corpus. Humaine Summer School, Genova, Italy, September 2006
- van Reekum, C. M., Johnstone, T., Banse, R., Etter, A., Wehrle, T., & Scherer, K. R. (2004) Psychophysiological responses to appraisal dimensions in a computer game. *Cognition and Emotion* 18, 663-688.
- Wang, N., & Marsella, S. (2006). Evg: an emotion evoking game. Proc. 6th International Conference on Intelligent Virtual Agents. Marina del Rey, CA, USA Berlin: Springer LNCS 282-291.
- Weizenbaum, J. 1996. ELIZA - A computer program for. the study of natural language communication between man and machine. *Comm. ACM* 9:36-45.